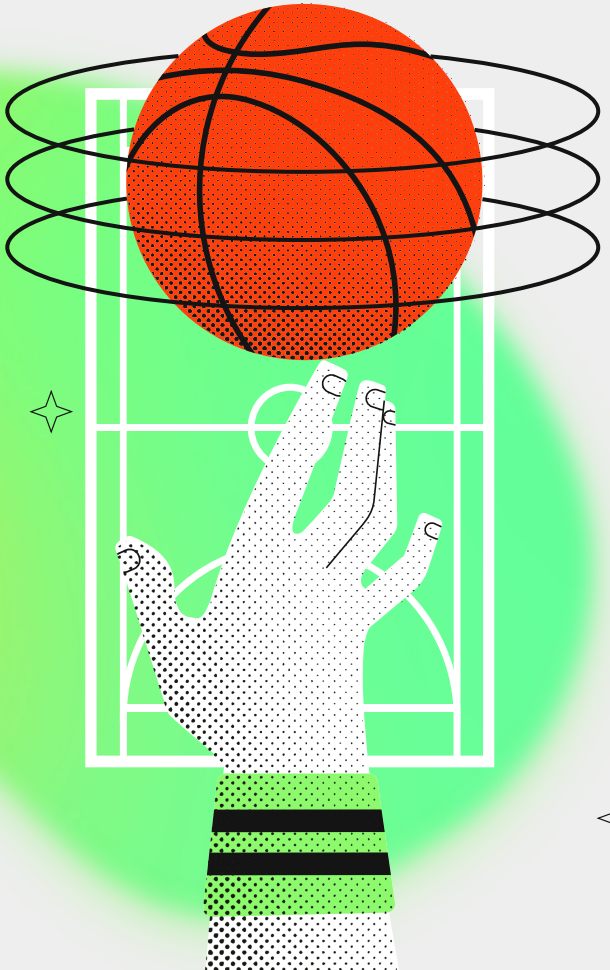
Predicting College Prospect Performance in the NBA

BA476 Spring 2023

Team 4: Cameron Anderegg, Anqi Chen, Beverly Liu, Vy Nguyen, Cecilia Zhang



Presentation Overview

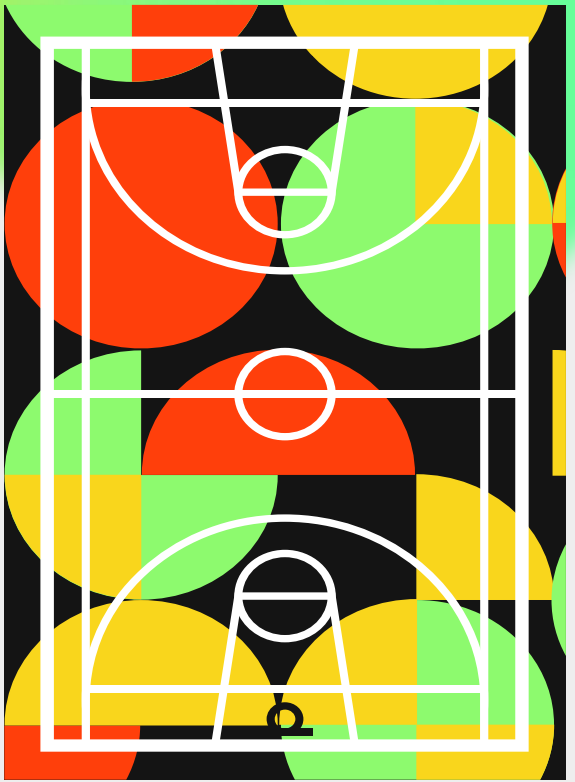
01 Project Objective

02 Data Collection

03 Model Overview

04 Model Discussion

05 Hypothetical Prediction



Can we predict the value of a college basketball player for their rookie NBA season?

- The NBA is a **highly competitive** league where teams **invest around $8 million per year** in scouting, signing and developing players every year.
- However, the process of evaluating players and making draft decisions is often **subjective** and can be influenced by personal **biases or preferences**, which can lead to **missed opportunities and poor performance**.
- By analyzing player data from their college careers, we aim to **predict how effective a draft pick will be in the NBA**. Our models can identify key predictors of NBA success and **make more accurate predictions**, potentially saving teams millions of dollars in player salaries and building stronger, more competitive rosters over time.

Data Preparation

Data we wanted

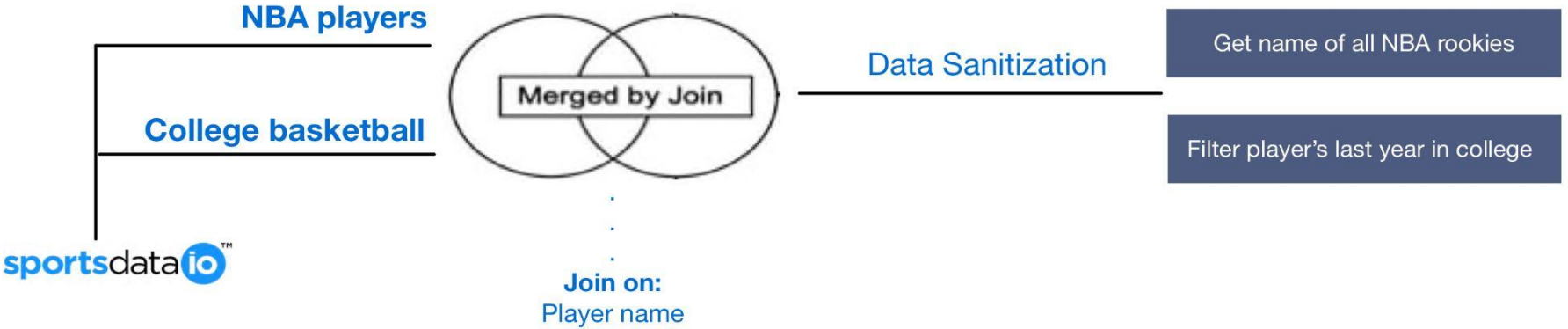
We wanted **normal** and **advanced** stats for a **college** players **final** season, as well as their advanced stats from their **NBA rookie** season

Kaggle vs API

We couldn't find a dataset from Kaggle so we picked an API called **SportsDataIO**. The reason we chose this instead of other API's is because this one was **free**, allowed for **multiple** API **subscriptions** and had helpful documentation.

Data from SportsDataIO

SportsDataIO had the data we wanted for both college players and NBA rookies **starting from the 2015-2016 season and beyond.**



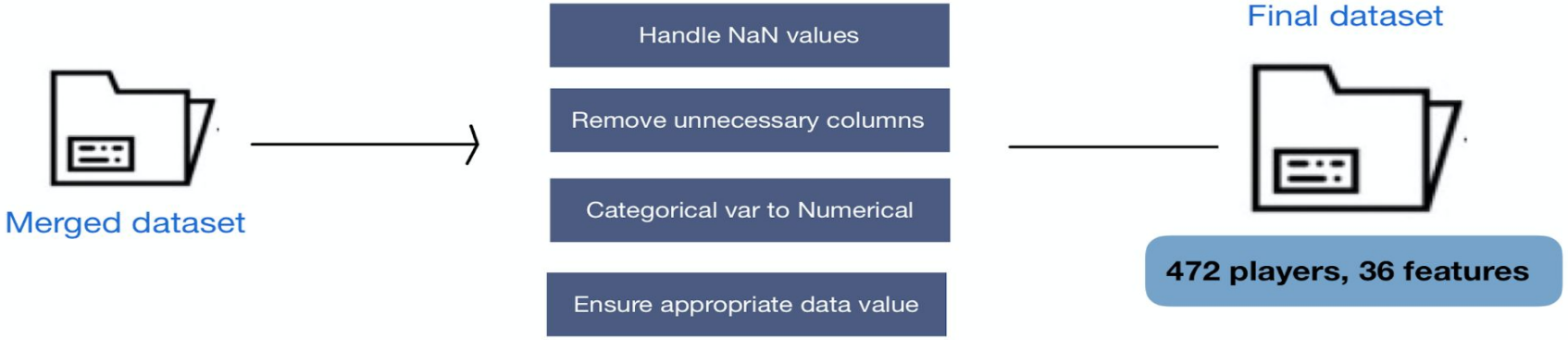
Data Cleaning

Data sanitization

Final dataset

One example of data **sanitization** is how we filtered the college players.
To find a player's **last year** in college, we would get the name of all NBA rookies (let's say for the 2019 class) then look for any **matching** names in the 2018 college season.

We decided we would have **one row** for **each player**, with the **columns** being all their **normal** and **advanced stats** from their last year of college, and an additional 2 columns with the year they were a rookie in the NBA and their player efficiency rating during this year.



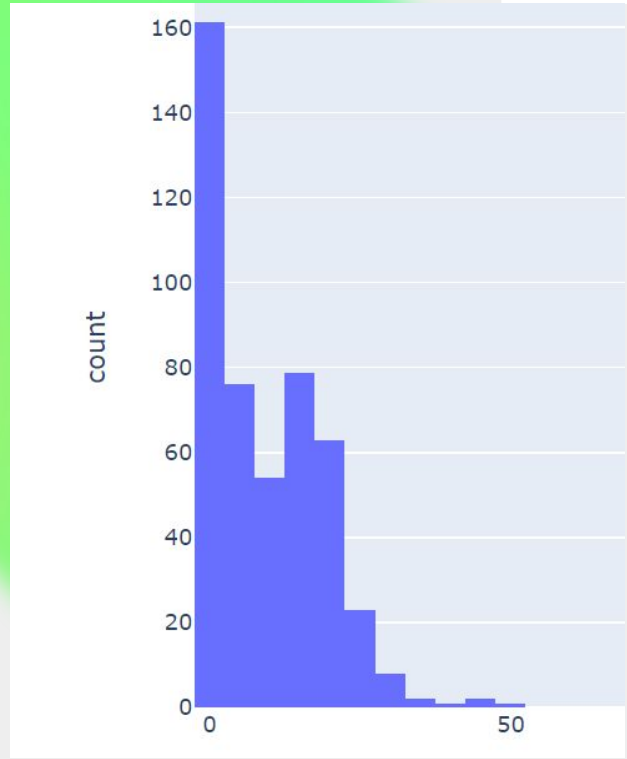
Target Variable Selection

Value

- What we really care about is value, but this is hard to measure (latent variable)
- Must proxy with something quantifiable

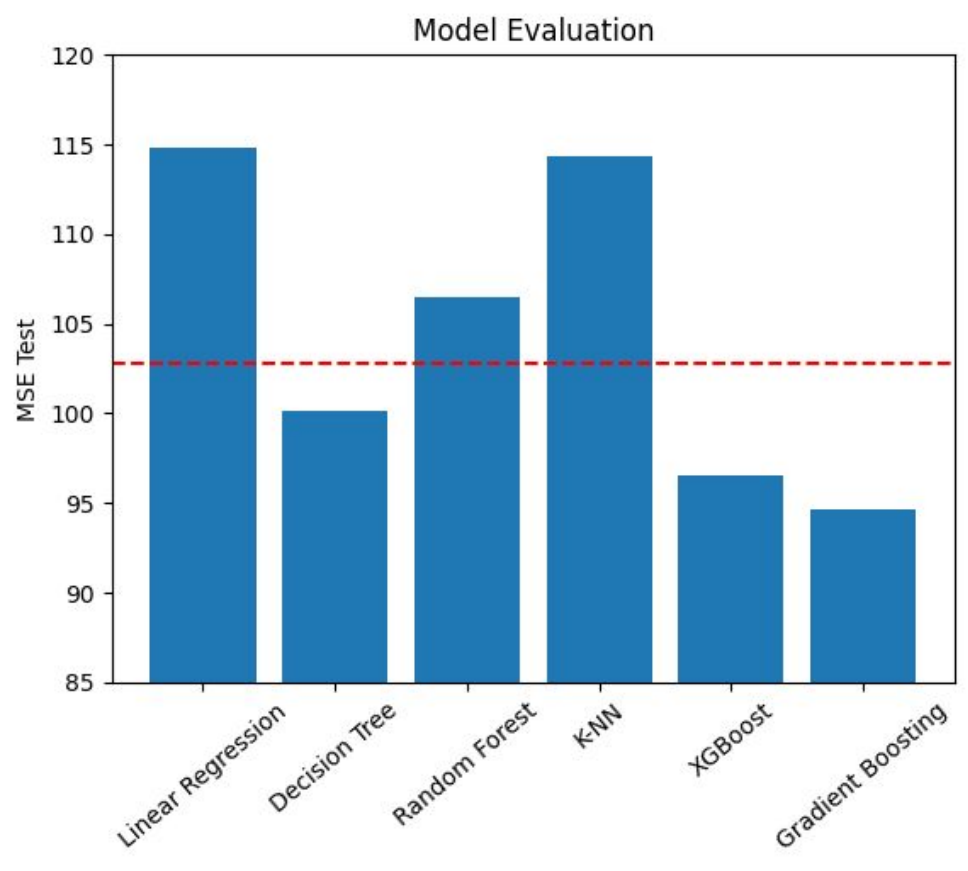
Player Efficiency Rating

- Proxy variable measures how effective a player is
- Relative to minutes to allow for comparison
- Understood that a rating above 35 = GOAT



Target variable distribution

Model Performance

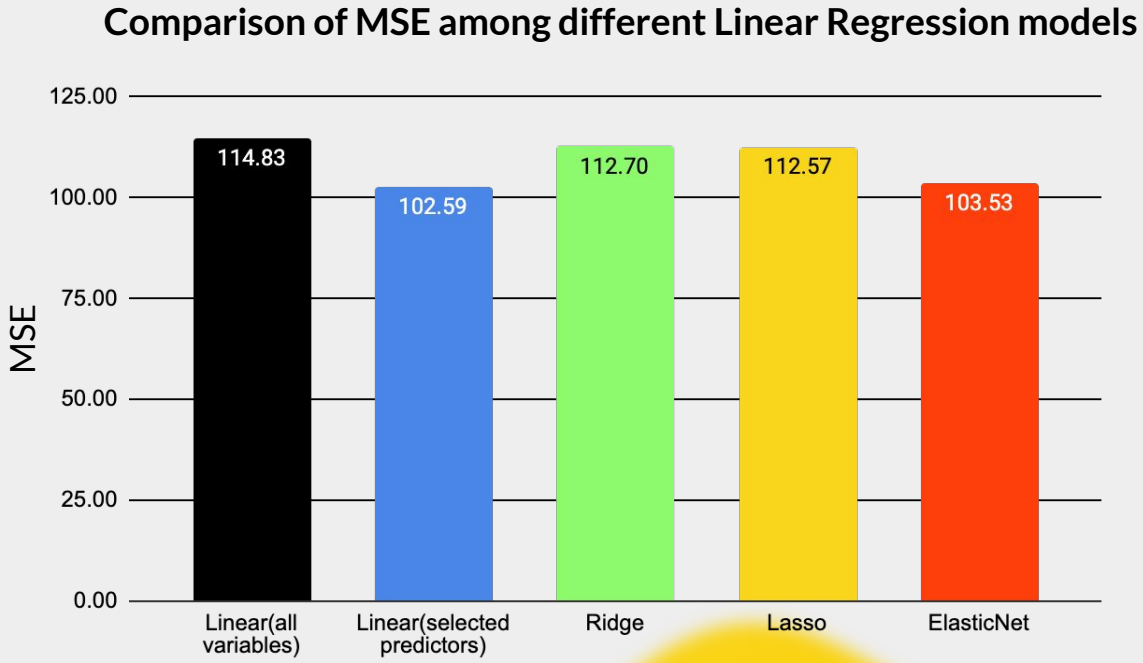


- Naive benchmark: predicting training average on test
- Models outperforming our naive model: decision tree, XGBoost and Gradient Boosting
- Models that did not outperform: Linear Regression, Random Forest, KNN
- Train test split based on season year cutoff

Linear Regression

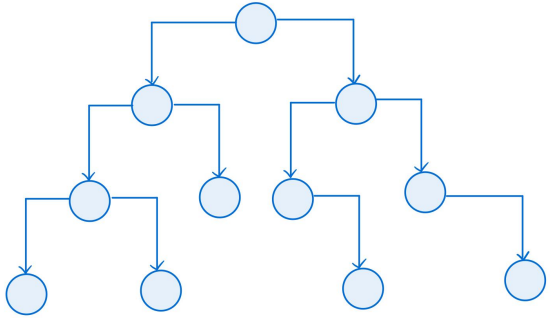
- The model with selected features (using Backward Elimination technique) has the MSE 102.59, which is lower than the benchmark. However, it may still not be effective to use for prediction as it contains multicollinearity.

- Among all ML Regularization methods, Elastic Net outperforms LASSO and RIDGE. It strengthens the point that using two penalties gives a much better result than using one.

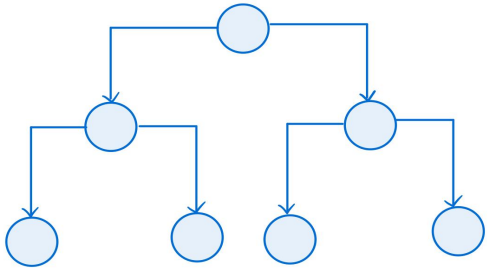


Tree-based Models

Construct 4 types of tree-based models: Decision Tree, Random Forest, XGBoost, Gradient Boosting Machine.

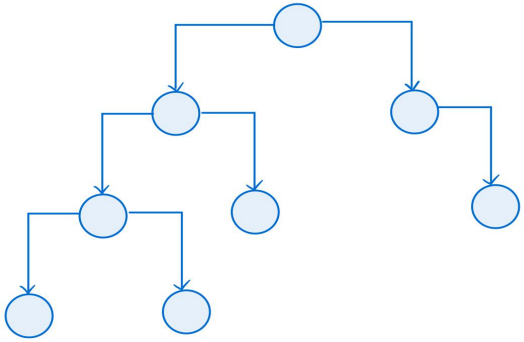


Full tree



Reduced tree

- Adding some parameters such as max_depth, learning-rates, n-estimators

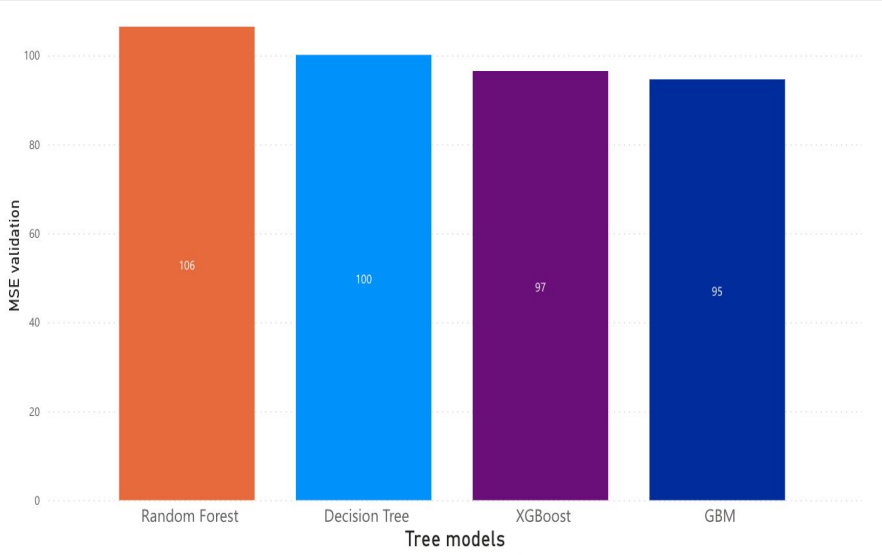


Tree with Hyperparameter Tuning

- Using GridSearchCV and RandomizedSearchCV.
- Selecting best values for parameters based on model's performance on validation set.

Boosting Ensemble works best

MSE Validation of each Tree model



- Ensemble method clearly shows better performance than even the best models of Decision Tree and Random Forest.
- Gradient Boosting gives off the best result. **Gradient Boosting Machine is our model of choice for prediction.**

Hypothetical Prediction

“March Madness”

Predicted Rookie Year PER: 12.33

- Popular opinion says this may be too low…
- Time will tell

Implications, Limitations and Future Avenues

Implications

- Evaluating the performance of current NBA players
- Evaluating draft prospects

Limitations

- Limited data quality/quantity may impact model accuracy.
- Unaccounted factors affect player performance and model predictions.
  - such as injuries, coaching, and team dynamics
- Model predictions not applicable to all players due to individual variability.
- Inability to capture tacit knowledge, with respect to NBA scouting

Future Avenues

- Updating the model with recent data to enhance accuracy and relevance.
- Applied to other professional basketball leagues or different sports to predict player performance.

Thank you!

Questions?